Amazon Co-Purchasing Optimization Report

Table of Contents

Introduction

On June 1, 2003, web crawling was conducted on the Amazon e-commerce site in order to create a network based on the *Customers Who Bought This Item Also Bought* feature on the site (Leskovec, Adamic, Huberman). After this initial data collection, Amazon contacted our team to perform an analysis of this data in order to optimize their current product offerings. In our initial project proposal, we highlighted that "on Amazon.com, somewhere between 20 to 40 percent of unit sales fall outside of its top 100,000 ranked products [Brynjolfsson et al. 2003]." Given this premise, our team of data scientists employed various visualization and modeling techniques to analyze network structures in the dataset in order to answer the overarching question: Can Amazon optimize their recommended products based on what products are frequently bought with other products in order to encourage more sales?

To conduct our analysis, we used the data based on the *Customers Who Bought This Item Also Bought* feature of the Amazon website that was collected on June 1, 2003. Within this dataset, a directed edge in the graph from node *i* to node *j* represented product *i* being frequently purchased with product *j.* The data set provided pairs of "From Node ID" corresponding to a "To Node ID" to indicate direct co-purchasing ties. One major limitation of this dataset, however, was that the names and the product identification numbers of these items were not provided. Without the actual names of these products, our team was limited in our ability to determine why certain products or types of products were linked or not linked, and we were limited in providing future direction on identifying specific optimization pathways to increase sales. However, one of the benefits of this directed network was our ability to analyze the direction of co-purchasing ties. In other words, we were able to visualize which items were directly purchased from and to

another item. This was especially helpful in investigating in-degree and out-degree relationships which may indicate which key products should be further promoted to increase sales based on their current popularity, in terms of being co-purchased with another product.

Furthermore, our initial dataset was extremely large given its substantial number of nodes. This dataset had 403,394 nodes and 3,387,388 edges. Due to the constraints of computational processing efficiency, our team strategically collected a subset of this data to conduct a more comprehensive and manageable analysis.

In order to select a subset of this data, our team thoroughly examined the entire dataset to identify a highly interconnected cluster that would best represent the Amazon co-purchasing network. We carefully evaluated the tradeoffs involved with extracting a segment of the data, and we placed a focused effort on achieving a balance between retaining network characteristics and enhancing processing efficiency.

Based on our preliminary analysis of the co-purchasing dataset, we identified 7 components in the network, and the largest strongly connected component contained 403,364 nodes. This component contained 98% of the total nodes and needed to be further reduced in size before proceeding to the community detection process. To do this, a random selection process was used, which is further detailed in the Analysis section of the report. Once the component was reduced to a quarter of its original size, we utilized the walktrap community detection algorithm to identify clusters within the data set. There were a total of 21,984 clusters identified, ranging in various sizes between 1 to 10850. Among these clusters, a highly interconnected cluster of 355 nodes and 812 edges was selected for our subset. Details of this subset data collection are further discussed in the Analysis section of this report.

Questions

Hence, to answer our overarching question of whether Amazon can optimize their recommended products based on which products are frequently bought with other products, we came up with some supporting questions to answer along the way. To begin with, we analyzed which products are frequently bought together, which can be important as Amazon could change their recommendation tactics or change how they sell certain products. For example, Amazon could consolidate certain products that are frequently bought together into a singular product bundle. We investigated this question through community detection, node-level metrics, and other analysis techniques as further discussed in the Analysis and Findings section of this report.

Additionally, we investigated whether there are certain products that result in longer-tailed connections or lead to more associated purchases. By identifying these products through centrality measures, such as eigenvector centrality, this information could help Amazon identify cross-selling opportunities and add more indirectly connected products to their recommendations feature to promote their sales.

Moreover, we analyzed which product was the most influential or led to purchasing the greatest number of other products. This could optimize Amazon's inventory management and supply chain logistics by recognizing which products are frequently co-purchased with other products. By identifying these key products using node-level metrics, such as degree centrality, Amazon could prepare ahead and streamline its stocking and delivery processes to increase cost savings.

Lastly, we examined whether purchasing patterns can be predicted, such as whether there was a tendency for item A to be bought with item B if item A is commonly bought with many other products, as well as a tendency for item A to be bought with item B if many items are

commonly purchased with item B. We utilized ERGM Modeling to conduct this predictive analysis which could aid in conducting product availability and customer behavior predictions for these recommended products.

Analysis

Data Collection

Given the substantial size of our preliminary dataset consisting of 403,393 nodes and 3,387,388 direct ties, we undertook the task of collecting a more manageable subset of this data for in-depth analysis. Our aim was to identify a highly interconnected cluster within the dataset that would best reflect the co-purchasing relationships of the Amazon items in the given dataset.

In order to select this subset of the data and visualize this network, we used the following packages in R: readr, tidytext, tidygraph, ggraph, igraph, tidyverse, topicmodels, textstem, udpipe, and stats.

Our team began by identifying the various components that exist within the entire network. We identified a total of 7 networks of the following sizes:

403364 , 2 , 2 , 3 , 3 , 15 , 5

The first component was significantly larger and more strongly connected compared to the other six components. This component had 403,364 nodes, which is 98% of the total nodes in the entire dataset. Due to its large size, we needed to scale down the size of the largest component to a quarter of its size before proceeding to community detection. To do this, we

utilized the set.seed() function and the sample() function to randomly select 100841 numbers between 1 to 403364. These numbers would then correspond to the nodes in our subgraph. There were several implications to conducting a random selection such as the introduction of a certain degree of bias and the possible omission of important nodes or connections. However, our ultimate objective was to strike a balance between retaining network characteristics and enhancing processing efficiency, and our team found that this approach would be most effective in doing so.

Once our dataset was more manageable in size, we would be able to utilize the walktrap community detection algorithm, cluster_walktrap(), in order to identify the clusters in the network. The walktrap algorithm was used because of its efficacy in handling large datasets and its compatibility with directed network structures, which matched the characteristics of our given dataset. It works by simulating random walks that explore node sequences, constructing a hierarchical clustering tree based on node similarities, and then cutting the tree to reveal distinct communities.

Upon conducting community detection, we identified 21,984 clusters in the subgraph with sizes from 1 to 10850. We selected one of these clusters with 355 nodes and 812 edges, which was ideal for the subset given that it ranges between 100 to 500 nodes. This cluster was used to form our subset for data analysis. Once the subset was identified, descriptive and predictive analysis was utilized. Figure 1 illustrates the original network of the selected subset.

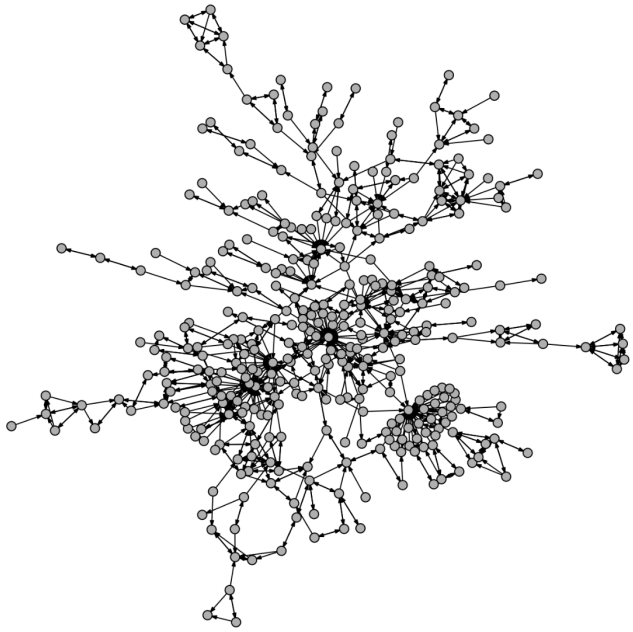


Figure 1. Subset Network Structure

Descriptive Analysis

Descriptive analysis of the data was conducted to better understand the network's structure. We looked at measures of centrality, the network's k-core decomposition, in-degree distribution, and out-degree distribution, and assessed whether the network had small world properties.

To conduct descriptive analysis, our team utilized the following packages in R: readr, tidytext, tidygraph, ggraph, igraph, tidyverse, topicmodels, textstem, udpipe, and stats. Because the subset we selected for our data analysis was initially a component of the entire network, it was unnecessary to extract the largest component of the network for further analysis, given the

subset itself was the only and largest component. Therefore, our analysis was conducted on the subset network as a whole.

Our team first began by investigating the local properties of the network. We computed various measures of centralities using the 'sna' package. These measures included in-degree, out-degree, betweenness, eigenvector, hub score, and authority score, which allowed us to see which specific nodes were the most influential based on these various centrality measures. By isolating nodes with high centrality, we would better understand which specific nodes or products to recommend or promote. Each centrality measure revealed various aspects of the data set. For example, degree centrality enabled us to identify incoming and outgoing connections between products, thus revealing which products may be a popular co-purchasing item. Betweenness centrality would be valuable in identifying which products may guide customers toward purchasing other products. The eigenvector centrality measure would allow us to locate products that are connected to frequently co-purchased products. On the other hand, hub scores would be helpful in identifying products that may point to other products with a lot of purchases, while authority scores may help to identify products that are frequently purchased.

Moreover, we used community detection algorithms to detect clusters within the network to identify products that are often purchased together. We utilized the walktrap algorithm as used in data collection due to its compatibility with directed graphs. This allowed us to identify the various communities or clusters among the co-purchased products.

In addition, we also calculated the k-core decomposition, and the in-degree and out-degree distribution to better understand the number of products customers purchase together. K-core decomposition would be helpful in identifying influential products within the network by continuously removing the least connected nodes from the network until the remaining nodes

have a minimum degree of k. This would reveal the densely connected core structures of the network. Furthermore, in-degree and out-degree distribution would indicate whether there are products that tend to be frequently bought together or exhibit patterns of one product leading to the purchase of another.

Additionally, we ran a clustering algorithm on our network to see how customers purchase products in groups and whether our network had small world properties to check whether customers purchased products in distinct categories. This analysis allowed us to gain a better understanding of the network structure and gain insights into how customers make purchases. By understanding pre-existing customer behavior and purchasing patterns, our aim was to optimize the recommendation system to improve the user experience and drive more sales.

Predictive Analysis - ERGM Model

In order to conduct predictive analysis of the network, our team utilized exponential random graph modeling (ERGM) techniques in R. The 'statnet' package was used to conduct ERGM-based network modeling to test the following hypotheses:

Hypothesis 1 (edges): Holding everything else constant for all, the probability that a tie exists between any two products will be very low.

Hypothesis 2 (mutual): Holding everything else constant for all, if item A is purchased with item B, it will be more likely that item B is purchased with item A.

Hypothesis 3 (odegree): Holding everything else constant for all, there is a tendency for item A to be bought with item B if item A is commonly bought with many other products.

Hypothesis 4 (idegree): Holding everything else constant for all, there is a tendency for item A to be bought with item B if many items are commonly purchased with item B.

These hypotheses would provide further insight to answer our main problem statement, "Can Amazon optimize their recommended products based on what products are frequently bought with other products in order to encourage more sales?" and enable us to predict purchasing patterns. For example, in Hypothesis 2, if we found that if a certain item A is purchased with item B, and it was likely that item B would be purchased with A, Amazon may consider creating bundles to encourage more sales. Furthermore, testing the indegree and outdegree distributions in Hypothesis 3 and 4 may help to identify the tendency for certain key products to be purchased frequently, which Amazon can leverage by further promoting these items.

To test these hypotheses, our team began by building the ERGM models. We specifically investigated endogenous statistics rather than exogenous statistics because we were not provided additional information on the Amazon items. Exogenous effects refer to effects of node attributes or variables outside the predicted ties. Rather, we used the following ERGM terms in R to observe endogenous statistics, which look at the ties with the subset network: edges, mutual, odegree, and idegree.

‘Edges’ was used in our ERGM modeling to test Hypothesis 1 as it allows us to examine the probability of a tie existing between two items. ‘Mutual’ was used to determine the probability of having pairs of reciprocated ties in order to test Hypothesis 2. In other words, if a customer who purchases item A purchases item B, the ‘mutual’ test would help to report the probability of a customer who purchases item B purchasing item A. ‘Odegree’ was used to investigate outdegree distribution which was relevant for Hypothesis 3 because it can be used to test whether there is a tendency for a small number of items to have a lot of co-purchased items. ‘Idegree” was used to test indegree distribution, specifically for Hypothesis 4 because it can be used to test whether there is a tendency for a small number of popular items to be purchased with another item. In other words, it can test whether there are a select few items that are popular co-purchased items. After collecting our results, we determined whether or not our hypotheses were supported given the p-value and the directionality of the effect. A p-value of less than 0.05 would indicate that the hypothesis is supported.

After testing our hypotheses with these four network statistics to conduct model estimation, we then proceeded to conduct model diagnostics. To do this, we used model convergence with the MCMC-MLE process (Markov Chain Monte Carlo - Maximum Likelihood Estimation) to test whether the estimation process converged to a desired state. Moreover, we used the goodness-of-fit test to determine whether our model estimate was a good representation of our subset data. P-values greater than 0.05 indicated a good fit.

Findings

Descriptive Analysis Findings

Since a node's eigenvector centrality indicates both its centrality and its proximity to influential nodes, we concluded that this would help us isolate products that are often bought along with other popular or high-demand items and are in high demand themselves. The items with the highest eigenvector centrality were Product ID 27656, 24718, and 27570; these products are both highly purchased and lead to the purchase of other highly purchased items. This measure helped us to identify key products that result in longer-tailed connections.

Other measures of centrality we looked at were a node's in-degree and betweenness. The in-degree of a node indicates how many incoming connections it has; therefore, a node with a high in-degree is likely a product to be purchased alongside a large number of products. According to our data analysis, the nodes with the highest in-degrees were Product ID 26776 with an indegree centrality score of 39, 21722 with 36, and 27565 with 30, suggesting that these would be good add-on products to recommend to customers. A node with high betweenness implies that the node acts as a bridge across different areas of the network. Nodes with high betweenness may guide customers to other different types of products and can increase the variety of products that customers are exposed to. The items with the highest betweenness centrality are 83673, 83672, and 43519.

We also looked at each node's hub and authority score. We concluded that nodes with high authority scores are products that are highly purchased, and products with elevated hub scores serve as pivotal points, connecting customers to a diverse range of reputable products. Although 'hub' products may not be directly purchased by customers, they play a vital role in

recommending products that customers are likely to be interested in. Nodes with high authority scores included Product IDs 27565, 24718, and 27570. Nodes with high hub scores include 326674, 143365, and 237829.

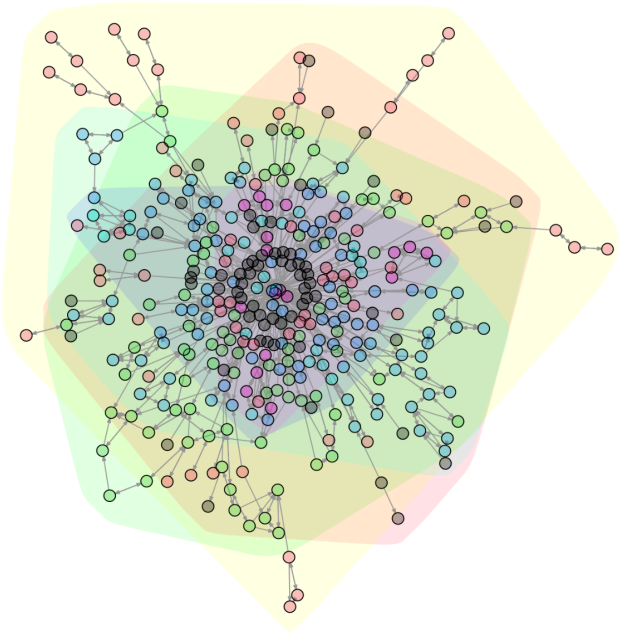


Figure 2. K-Core Decomposition Plot

In addition, we looked at the k-core decomposition plot and k-core scores in order to learn more about the network structure as well as its connectivity and density. Higher k-core scores are indicative of greater influence within the network, often corresponding to nodes that hold more central positions. The maximum k-core in our data was 6, suggesting that a subgraph exists where each node is connected to at least 6 other nodes within that subgraph. Some nodes with a degree of 6 included: 16770, 16771, 35897, and 35898, suggesting they are highly influential and should be strongly marketed to customers.

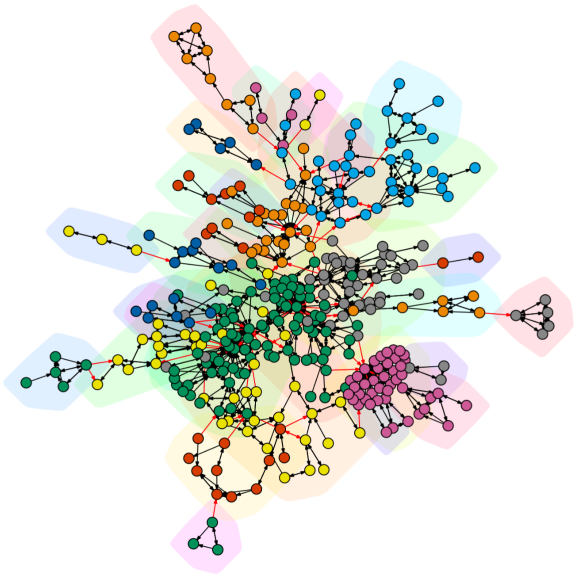


Figure 3. Community Structure

Using the walktrap algorithm, we calculated a modality score of 0.8198 and 32 clusters present in our network. The high modularity score indicates a well-defined community structure. Each node in each cluster is well-connected to another in comparison to nodes outside of the cluster. Additionally, this clustering showed that there are 32 somewhat distinct groups of products, which can be used to inform how products are grouped together on the Amazon homepage. Cluster sizes ranged from 1 to 45 as shown in Figure 4; since there are some clusters that are smaller in size, it is possible that a miscellaneous category should be made or certain categories may need to be grouped together. Additionally, there were some clusters that only consisted of one or two nodes, suggesting that some products are purchased alone or in very small groups.

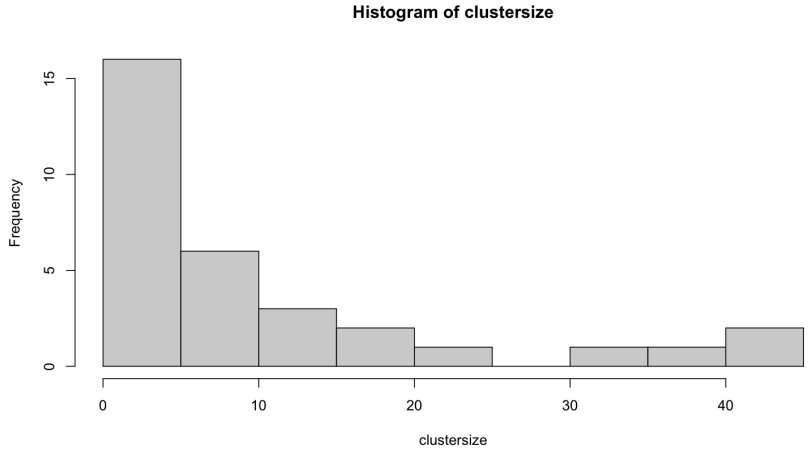


Figure 4. Histogram of Cluster Size

c

Figure 5. In-degree Distribution

The indegree distribution plot in Figure 5 suggested that there are a significant number of nodes with a low indegree of less than 5. The distribution of indegrees ranging from 10-40 was relatively similar; however, the number of nodes with indegrees greater than 10 was quite low. This analysis suggests that most products are commonly bought with a small number of other products, and so Amazon recommendations may focus on these key connections rather than more distantly related products.

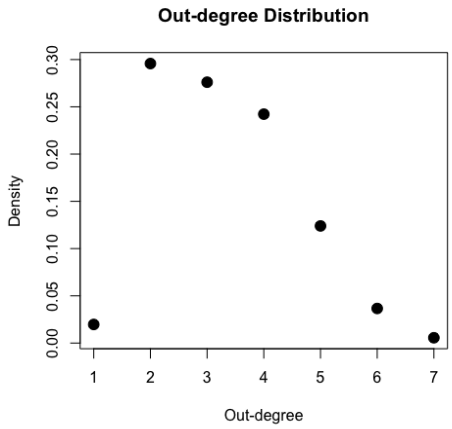


Figure 6. Out-degree Distribution

The outdegree distribution plot in Figure 6 peaked at 2 suggesting that there is a significant number of products that lead to the purchases of two other items. The outdegree peak of 2 may suggest that items are commonly bought in groups of 3. This may explain why Amazon recommends "Frequently bought together" items in groups of 3. Outdegrees of 3-4 are also quite common, although the frequencies decrease as the outdegree increases.

Our network also demonstrated small world properties due to a higher clustering coefficient and a shorter average path length than expected by random chance, suggesting that customers can quickly navigate to and from products that are further in the network. In other words, most products' neighbors are also neighbors with each other. This was an interesting finding given the above findings that most products have low degree, likely meaning that related products tend to form dense clusters in our network.

Predictive Analysis Findings - ERGM Model

Next, we investigated the structural network relationships through an ERGM model, used for predictive modeling. As detailed above, we created an ERGM model on the following network statistics: edges, mutual, gwodegree, and gwidegree. Table 1 shows the model summary,

including parameter estimates for each of the model arguments. Each of the parameters was highly significant given that their p-values were less than 0.05, and these findings led us to conclusive evidence for our hypotheses.

```
> summary(model1)
Call:
ergm(formula = item ~ edges + mutual + gwodegree(log(2), fixed = T) +
    gwidegree(log(2), fixed = T))

Monte Carlo Maximum Likelihood Results:

                              Estimate Std. Error MCMC % z value Pr(>|z|)
edges                          -5.9002     0.1104      0 -53.420   <1e-04 ***
mutual                          6.4791     0.1524      0  42.526   <1e-04 ***
gwodeg.fixed.0.693147180559945  2.6619     0.2892      0   9.203   <1e-04 ***
gwideg.fixed.0.693147180559945 -2.8390     0.1365      0 -20.805   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 174216  on 125670  degrees of freedom
 Residual Deviance:   7702  on 125666  degrees of freedom

AIC: 7710  BIC: 7749  (Smaller is better. MC Std. Err. = 3.086)
```

Table 1. Summary of ERGM Model Results

Hypothesis 1 (edges): Holding everything else constant, the probability that a tie exists between any two products will be very low. In other words, the network will not be very dense.

Result: The model supports the hypothesis. Since the edges parameter estimate is -5.9002, the probability that a tie exists is 0.005. In other words, holding everything constant, the probability that a tie exists between item A and B is 0.5%. This corresponds closely to the density of the network, 0.006, and represents a very sparse network of connections.

Hypothesis 2 (mutual): Holding everything else constant, if item A is purchased with item B, it will be more likely that item B is purchased with item A.

Result: The model supports the hypothesis. As evidenced by the strong positive estimate of the parameter, ‘mutual’, holding everything else constant for all, if item A is purchased with item B,

it will be more likely that item B is purchased with item A. Converting from the log-odds ratio of 6.4791 to a probability, we find that the probability that if item A is co-purchased with B, that item B will also be co-purchased with item A is 99.8%.

Hypothesis 3 (gwodeg): Holding everything else constant for all, there is a tendency for item A to be bought with item B if item A is commonly bought with many other products.
Result: The model contradicts the hypothesis. As evidenced by the strong positive estimate of the parameter for gwodeg, holding everything else constant, there is a tendency against outgoing ties originating from nodes that already have other outgoing ties. Converting from the log-odds ratio of 2.6619 to a probability, we find that the probability that outgoing ties are not directed towards nodes with other outgoing ties is 93.5%.

Hypothesis 4 (gwideg): Holding everything else constant for all, there is a tendency for item A to be bought with item B if many items are commonly purchased with item B.
Result: The model supports the hypothesis. As evidenced by the strong negative estimate of the parameter for gwideg, holding everything else constant, incoming ties are more likely to be directed towards nodes that already have other incoming ties. In other words, there is a tendency for item A to be bought with item B if many items are commonly purchased with item A. Converting from the log-odds ratio of -2.8390 to a probability, we find that the probability that incoming ties are not directed toward nodes that already have many incoming ties is 5.5%.

We investigated the quality of the ERGM model and its conclusions through goodness of fit testing. First, we determined that the model converged. Divergence would indicate that the

model is unusable and unstable, even if the parameter estimates have statistical significance. The trace plots were relatively stable around 0 and did not exhibit any nonlinear patterns, as shown in Figure 6. The distributions of the sample statistic deviations were fairly normally distributed and unimodal with a mean of 0. Therefore, we were able to move forward with this analysis.

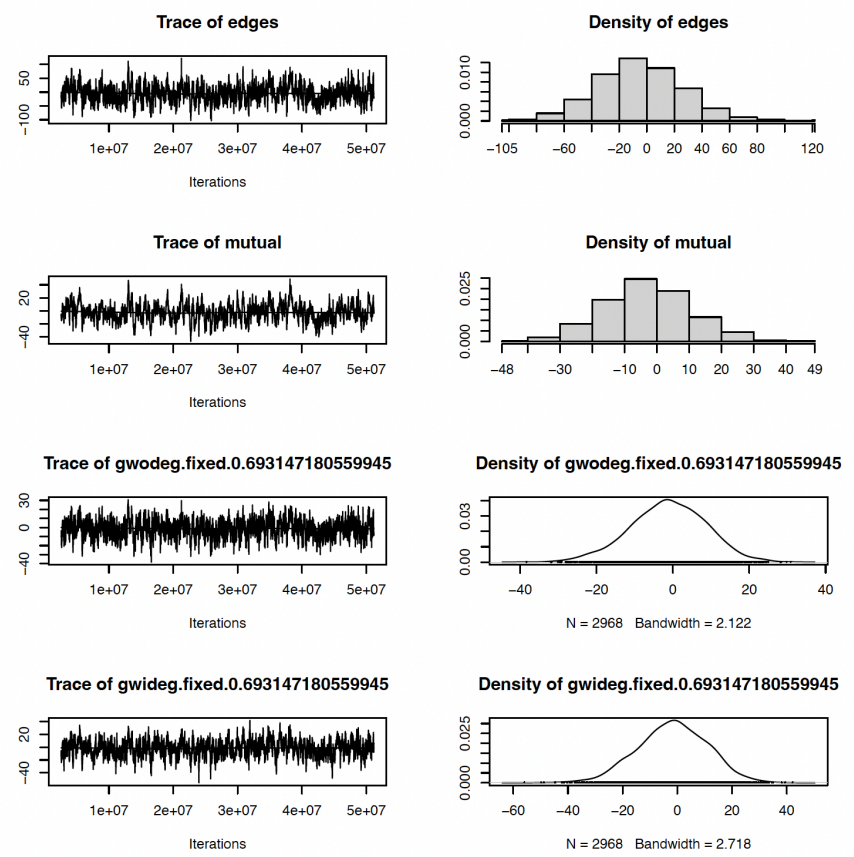


Figure 6: Evaluation of Convergence

Secondly, we examined the goodness of fit through the gof function available in the ergm package in R. Table 2 shows the results of the analysis. Most of the p-values for outdegree were greater than 0.05, so the model fit well with the original network in terms of the outdegree distribution. However, most of the p-values for indegree were less than 0.05, so we were unable to conclude that the model is a good fit in terms of indegree distribution. In other words, the

model did not capture the indegree distribution of the original network well. Therefore, we cannot make a strong conclusion about Hypothesis 4, but rather make more general statements about the direction of the parameter estimate and whether it makes sense. Further results of the goodness-of-fit testing are available in Appendix A.

Goodness-of-fit for in-degree

	obs	min	mean	max	MC p-value
idegree0	91	99	120.420	140	0.00
idegree1	103	36	54.590	77	0.00
idegree2	63	25	41.915	63	0.01
idegree3	37	19	37.590	60	0.94
idegree4	26	13	32.605	49	0.24
idegree5	9	17	27.240	40	0.00
idegree6	9	9	18.525	31	0.02
idegree7	2	4	11.530	20	0.00
idegree8	0	0	6.085	14	0.01
idegree9	3	0	2.850	8	1.00
idegree10	3	0	1.070	5	0.18
idegree11	0	0	0.425	2	1.00
idegree12	1	0	0.115	2	0.22
idegree13	2	0	0.035	1	0.00
idegree15	1	0	0.005	1	0.01
idegree17	1	0	0.000	0	0.00
idegree23	1	0	0.000	0	0.00
idegree30	1	0	0.000	0	0.00
idegree36	1	0	0.000	0	0.00
idegree39	1	0	0.000	0	0.00

Goodness-of-fit for out-degree

	obs	min	mean	max	MC p-value
odegree0	7	3	10.695	22	0.35
odegree1	105	68	90.385	113	0.06
odegree2	98	103	124.920	152	0.00
odegree3	86	56	77.995	100	0.35
odegree4	44	18	34.195	52	0.10
odegree5	13	2	12.025	22	0.87
odegree6	2	0	3.625	9	0.61
odegree7	0	0	0.920	6	0.91
odegree8	0	0	0.190	2	1.00
odegree9	0	0	0.045	1	1.00
odegree11	0	0	0.005	1	1.00

Goodness-of-fit for edgewise shared partner

	obs	min	mean	max	MC p-value
esp.OTP0	388	621	776.245	838	0
esp.OTP1	256	2	22.905	112	0
esp.OTP2	128	0	1.900	41	0
esp.OTP3	37	0	0.235	7	0
esp.OTP4	3	0	0.000	0	0

Goodness-of-fit for minimum geodesic distance

	obs	min	mean	max	MC p-value
1	812	723	801.285	849	0.71
2	948	1580	1970.190	2280	0.00
3	710	2780	4363.630	5666	0.00
4	563	4710	8549.040	11612	0.00
5	442	7058	13451.595	17212	0.00
6	343	9168	15536.630	18356	0.00
7	272	9803	12868.525	15529	0.00
8	199	4769	8010.635	11483	0.00
9	153	1516	4058.235	8506	0.00
10	112	312	1796.530	6104	0.00
11	111	33	725.960	3946	0.10
12	44	0	276.900	2433	0.26
13	33	0	103.165	1443	0.87
14	20	0	38.635	878	0.63
15	0	0	14.375	482	0.96
16	0	0	5.440	254	1.00
17	0	0	2.005	126	1.00
18	0	0	0.765	76	1.00
19	0	0	0.310	42	1.00
20	0	0	0.155	25	1.00
21	0	0	0.055	10	1.00
22	0	0	0.010	2	1.00
23	0	0	0.005	1	1.00
Inf	120908	45026	53095.925	62107	0.00

Goodness-of-fit for model statistics

	obs	min	mean	max	MC p-value
edges	812.0000	723.0000	801.2850	849.0000	0.71
mutual	212.0000	184.0000	206.7050	227.0000	0.60
gwodeg.fixed.0.693147180559945	514.1250	484.3438	511.1110	530.7031	0.78
gwideg.fixed.0.693147180559945	380.1064	345.3550	377.6166	403.5508	0.84

Table 2. Goodness-of-fit Test Results

The ERGM model led us to several key conclusions about the Amazon co-purchasing network, regarding its physical characteristics, structure, and the nature of relationships. First, holding everything constant, the probability that a tie exists between item A and B was 0.5%. This conclusion was not surprising, since we expected the network to be very sparse. Between any two random items, we would not expect there to be a relationship. Next, the model showed that if item A is purchased with item B, it is extremely likely that item B is also commonly purchased with item A, demonstrating the highly mutualistic nature of the purchases. This also agreed with our intuition, because we would expect co-purchasing ties to go both ways. The most surprising finding was the conclusion of Hypothesis 3, that there is a tendency against

outgoing ties originating from nodes that already have other outgoing ties. However, on the other hand, while it was not conclusive, Hypothesis 4 indicated that incoming ties would be more likely to be directed towards nodes that already have other incoming ties. This difference is explained by the directionality of the links — what it means to be commonly bought with other items, or to be the item that is commonly bought with other items. For one, there are no nodes with an outdegree greater than 7, while the indegree plot, in Figure 3, shows a steep slope, with a long right skewed tail, displaying that there are nodes with an indegree of over 40. This means that there are a few nodes that have many incoming links, or are products that are purchased commonly with many others, but no nodes have a very high outdegree. This might be explained by the fact that there are only so many products that a single item might be connected to via an outward tie. However, a product can have a much larger number of incoming ties due to the nature of hubs and popular items on the Amazon website.

Implications

According to our findings, products with high centrality scores play a critical role in the network, either as connectors or influential items. They can also drive the purchase of other popular items and significantly impact sales. We suggest that Amazon should focus marketing efforts on products with high eigenvector centrality. These products are not only popular on their own but also have the potential to guide customers to other products, increasing their overall exposure to the product catalog. They should be suggested on the homepage or at the top of a user search.

In addition, Amazon should use products with high in-degrees and betweenness centrality for add-on recommendations. When customers view a product, suggest other items that frequently accompany it in customers' purchases, thereby encouraging larger orders and exploration of more diverse products. Furthermore, Amazon could also try an “Amazon’s Picks” labeling system and prioritize products with high degree centrality. Customers could be swayed to click on these products labeled as such which in turn could lead them down the line of products that are commonly bought along that “Amazon Pick” product.

While products with elevated hub-scores may not directly result in immediate purchases, their strategic placement can also guide customers towards a number of items with high authority scores that are more likely to be purchased.

The presence of well-defined clusters in the network suggests that products naturally group together based on customer purchasing patterns. Organizing products on the Amazon homepage according to these clusters could improve navigation and make it more intuitive for customers to discover related products, increasing sales. Since some clusters are smaller, they may need to be combined with larger clusters or grouped together as miscellaneous. Since 2003, when this dataset was taken, to the present, Amazon has utilized a system of tabs at the very top of its website. Back in 2003, the tabs were of the handful of categories of items that Amazon offered like books, electronics, and DVDs. However, with Amazon's ever growing catalog of items, the tabs at the top of the page are now being utilized for things like special deals, the AmazonBasics brand, and a rotating selection of items pertaining to a certain time of year like back to school or holidays. These rotating tabs are even more important for website navigation and could be a great place to put categories of the items we have found to be a part of the larger more well-defined clusters within the network

Given the network's characteristics, we also believe that Amazon should refine the recommendation algorithms based on existing customer purchasing patterns. They should reduce the emphasis on recommending products with low connectivity, as they may be less likely to lead to purchases or exploration of new products. They can also utilize the small world properties of the network. Since customers are likely to make smaller purchases of closely related items, Amazon should suggest related items in a cluster instead of distant products. However, by suggesting nodes with high betweenness, Amazon can also allow customers to explore a diverse range of products and navigate between dense clusters.

We also recommend that Amazon should use the insights gained from the out-degree distribution to design customer engagement strategies. For example, they should continue to focus on promoting products that tend to be bought together in groups of three, as indicated by the peak. They could also track purchasing patterns of commonly bought items, and create bundled deals or product packages; this would provide customers with convenient options of items that are commonly purchased together and increase sales.

Moreover, with our findings as a jumping off point, there are additional avenues and further steps forward that Amazon could take: further network research, discontinuing or finding better items to pair with isolated products, exploring trends within clusters, and product bundling. Without knowing what products each node is, however, makes it difficult to make concrete specifications since we can only make assumptions.

Bundling products is another great opportunity to utilize what we have observed about out-degree distribution within the network. Amazon’s Prime Day would be a very opportune time to offer exclusive bundled packages on discount containing the products that are the most often bought together. A customer might log onto Prime Day and see that the products that they

are always buying together are now in a Prime Day exclusive discounted bundle and take the opportunity to stock up on their favorites.

A highly valuable insight would be the type of products that are central, why they might be so central, and why they lead to the purchase of more products. Are they items that can't be used alone? What good are dish gloves without a dish sponge? What good is a dish sponge without dish soap? If you've got a dish sponge maybe you will now need a little sponge holder for your sink! Understanding the origin of a chain reaction of purchases like this can be helpful in pushing these kinds of products to the front of Amazon's product catalog. We encourage Amazon to utilize the available metadata on products to understand the network and its relationships. The ERGM, when used on a network with exogenous statistics, can be particularly useful in understanding why networks are the way they are and further explaining purchasing behavior.

There are many subsequent steps that could be taken by Amazon to further delve into the questions and main problem, but the biggest limitation within our own analysis was the limited data we had to work with. We scaled down the size of the dataset in order to make it usable with the limited computing resources we have, our laptops. Amazon itself is a large cloud computing provider, through Amazon Web Services, and can run graphical networks and machine learning models on an immense scale. We recommend investing in these large models to understand the more than 12 million products that exist on the marketplace. With a larger and more specific database, Amazon can specifically pinpoint the products in this network that will encourage co-purchasing. Additionally, through its pre-existing advanced machine learning and artificial intelligence capabilities, Amazon can truly harness the full power of the co-purchasing data. Machine learning models, trained on the network data, could be used to predict what product a

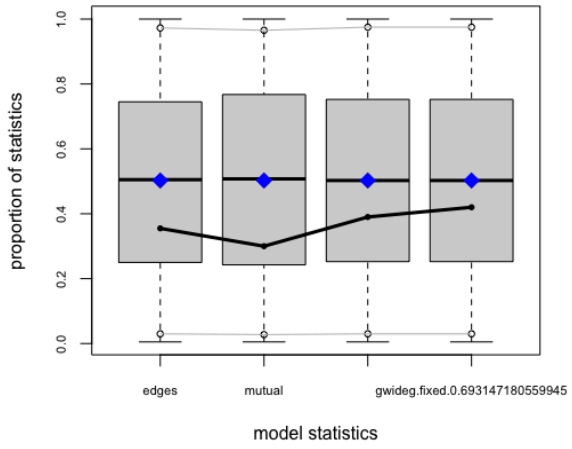
customer is most likely to purchase next, given their purchasing and even recent viewing or clickstream history. For more pertinent analysis, we would recommend that Amazon take a similar dataset from the past year or two to get an updated understanding of what their network now looks like. Alternatively, Amazon could incorporate real-time data into a graphical model in order to produce recommendations that update instantaneously. Since 2003, Amazon has expanded globally and the products frequently bought together in the United States are most likely very different from the global market. Rather than looking at all international purchases as one network, breaking each network down by country could lead to more regionally targeted recommendations.

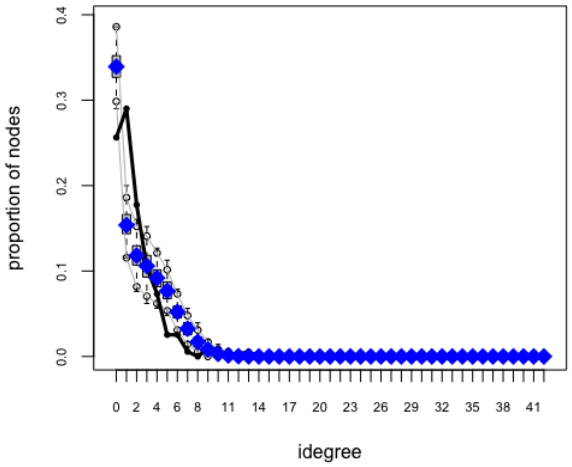
Reflection

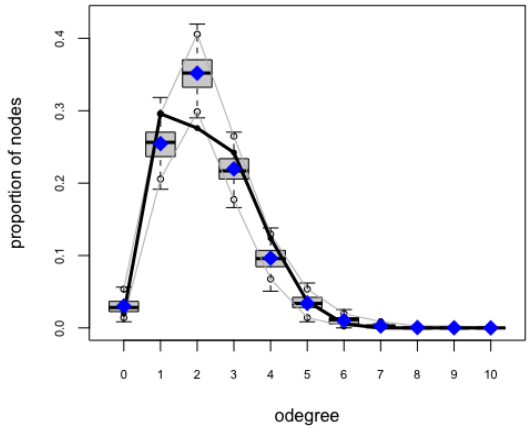
Overall, our data analysis yielded us the results we needed to answer the questions we posed to solve our overall problem statement. A question we posed for our ERGM model was: is there a tendency for item A to be bought with item B if item A is commonly bought with many other products? This was one hypothesis we were not able to draw a strong conclusion upon due to an inadequate p-value for the goodness of fit. Although it would have been good to be able to reach a conclusion on this hypothesis, it did not hinder our ability to answer the question we posed for the model since we had about transitive relationships in our network since we had other hypotheses we were able to draw conclusions on. Other than this, the work we have done with community detection, node-level metrics, and network visualizations was adequate for us to come to conclusions for our proposed questions. In general the predictions we discussed were reflected in the results and analysis. There was nothing particularly surprising or unexpected about our results.

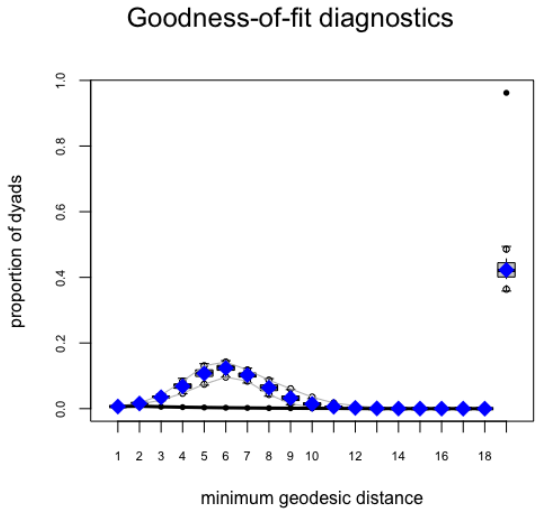
Appendices

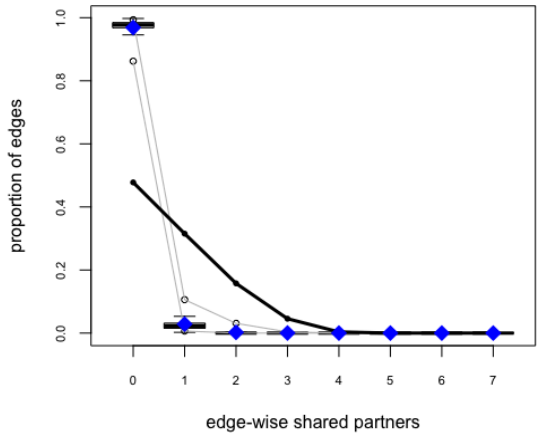
Appendix A. ERGM Goodness-of-fit Plots










proportion of edges
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
edge-wise shared partners